DOCUMENT RESUME

ED 227 146 TM 830 152

AUTHOR Ironson, Gail H.; Craig, Robert
TITLE Item Bias Techniques When Amount of Bias Is Varied and Score Differences between Groups Are Present. Final Report.
INSTITUTION University of South Florida, Tampa. Dept. of Psychology.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE 82
GRANT NIE-G-81-0045
NOTE 45p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability Grouping; Difficulty Level; Higher Education; *Item Analysis; Latent Trait Theory; Scores; *Sex Differences; *Test Bias; Testing Problems; *Test Items; Test Reliability
IDENTIFIERS Chi Square; Item Parameters

ABSTRACT

This study was designed to increase knowledge of the functioning of item bias techniques in detecting biased items. Previous studies have used computer-generated data or real data with unknown amounts of bias. The present project extends previous studies by using items that are logically generated and subjectively evaluated a priori to be biased or unbiased, and simultaneously controls the amount of bias and true ability differences (as measured by the unbiased items). The study evaluated the functioning of four statistical methods of assessing test item bias (transformed item difficulties, chi-square, three parameter and one parameter item characteristic curves) when (1) tests have varying amounts of bias (0 biased/60 items, 18 biased/78 items, 40 biased/100 items) and (2) ability differences on the unbiased items were either one-half or one standard deviation apart. Results indicate that agreement among the methods and between the statistical methods and judged bias was generally high except for data set VI (40 percent biased items, one standard deviation difference). Problems with individual methods and with cutoffs are discussed. Finally, presence of biased items did not affect reliability but did decrease validity and increase score differences between groups. (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227146

FINAL REPORT

Grant NIE-G-81-0045

ITEM BIAS TECHNIQUES WHEN AMOUNT OF BIAS IS VARIED AND SCORE DIFFERENCES BETWEEN GROUPS ARE PRESENT

Principal Investigator

Gail H. Ironson, PhD

Research Assistant

Robert Craig, MA

Department of Psychology
University of South Florida
Spring 1982

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

This research was sponsored with funds from the National Institute of Education, Grant NIE-G-81-0045. The views and opinions expressed in it are those of the authors and do not necessarily reflect those of the National Institute of Education.

Special thanks to Lawrence Rudner, PhD, who was the Project Director for this Grant.

TM 830 152

TABLE OF CONTENTS

	Page
ABSTRACT	iii
LIST OF TABLES	iv
INTRODUCTION	1
PURPOSE	2
REVIEW OF ITEM BIAS METHODS	5
Example of a Biased Item	5
Transformed Item Difficulties (TID)	6
Chi-Square (CHI)	7
Three Parameter Item Characteristic Curve	9
One Parameter Item Characteristic Curve	10
PROCEDURES	10
Sample	10
Research Instrument Development	12
1. Percentage of biased items	13
2. Observed differences in score distributions of males and females	14
Six Data Sets Used	16
RESULTS	18
Bias Methods Used in the Analysis	18
Agreement Among Bias Methods	22
Agreement Between Subjective Methods of Bias and Statistical Methods	23
Agreement Between Statistical Methods and Biased/Unbiased Classification	25
Psychometric Properties of the Tests	27
CONCLUSION	33

ABSTRACT

This study was designed to increase knowledge of the functioning of item bias techniques in detecting biased items. Previous studies have used computer-generated data or real data with unknown amounts of bias. The present project extends previous studies by using items that are logically generated and subjectively evaluated a priori to be biased or unbiased, and simultaneously controls the amount of bias and true ability differences (as measured by the unbiased items).

The study evaluated the functioning of four statistical methods of assessing test item bias (transformed item difficulties, chi-square, three parameter and one parameter item characteristic curves) when (1) tests have varying amounts of bias (0/60 items, 18 biased/78 items, 40 biased/100 items) and (2) ability differences on the unbiased items were either one half or one standard deviation apart.

Results indicate that agreement among the methods and between the statistical methods and judged bias was generally high except for data set VI (40% biased items, one standard deviation difference). Problems with individual methods and with cutoffs are discussed. Finally, presence of biased items did not affect reliability but did decrease validity and increase score differences between groups.

LIST OF TABLES

Number		Page
1.	Characteristics of Data Sets Used in this Study	17
2.	Means and Standard Deviations of Bias Methods	20
3.	Intercorrelations Among Signed Bias Methods	21
4.	Correlations of Signed Bias Indices with Judged Bias	24
5.	Identification of Biased Items Using Selected Cutoffs for Statistical Procedures	26
6.	Observed Means and Standard Deviations for Six Conditions	28
7.	Reliabilities of Tests Composed of Varying Amounts of Bias	29
8.	Factor Analysis Results	30
9.	Validities of Tests Composed of Varying Amounts of Bias	31

ITEM BIAS TECHNIQUES WHEN AMOUNT OF BIAS IS VARIED AND SCORE DIFFERENCES BETWEEN GROUPS ARE PRESENT

The issue of bias in measurement and selection is an important one in allowing equal opportunity for persons of equal ability regardless of whatever disadvantaged group to which they may belong. Since tests are increasingly used as devices for evaluation and placement, test constructors must make every effort to remove bias from them. Minority groups claim that traditional education and employment tests may not be measuring their true ability since the tests are based on the cultural experiences of the white middle class (Williams, 1970, 1971).

In recognition of this problem, two conferences (National Institute of Education, 1975; U.S. Office of Education, 1976) were held addressing questions of bias. In addition, sessions at several national organizations (AERA, APA, NCME) were devoted to examining items for bias in 1978 and 1979. The literature has proliferated in the last few years (for example, JEM, 1976) and has followed two main streams of inquiry: (1) Bias in selection covering predictions made by a test in the presence of an external criterion; (2) item bias studied in the absence of an external criterion (which would be most useful during test development).

The present study attempts to address some questions not yet explored by recent research in the area of item bias. Previous studies have used computer generated data or real data with unknown amounts of bias. This study proposes to find out which method is best under a

variety of conditions aimed at simulating various features of realistic conditions. This information is essential because the methods differ widely in terms of cost, sample size required, and ease of implementation.

PURPOSE

The literature on item bias contains several excellent review (Merz, 1978; Peterson, 1977; Rudner, Getson, & Knight, 1980). Various methods that have been explored include: (1) Analysis of Variance (Cardall & Coffman, 1964; Cleary & Hilton, 1968); (2) transformed item difficulties (Angoff & Ford, 1973); (3) discrimination measures (Green & Draper, 1972; Ozenne, Van Gelder, & Cohen, 1974); (4) item characteristic curves; (Ironson, 1982; Lord, 1977; Wright, Mead, & Draba, 1976); (5) chi-square (Scheuneman, 1979); (6) multivariate factor structures (Green, 1976; Green & Draper, 1972; Merz, 1973, 1976a); (7) response foil approach (Veale & Foreman, 1976).

Recent research has attempted to examine how effectively these methods identify biased items and the concordance among the methods. Ironson and Subkoviak (1979) found support for the three parameter item characteristic curve approach, the chi-square procedure, and the transformed item difficulty procedure. Rudner, Getson, and Knight (1979) found most support for the three parameter procedure, a chi-square procedure using five intervals, and a transformed difficulty approach. Merz and Grossen (1979) favored the transformed item difficulties procedure.

These recent studies on item bias as well as previous ones have either used existing data sets where the amount of bias is unknown a priori (Ironson & Subkoviak, 1979; Nungester, 1977; Rudner & Convey, 1978; Scheuneman, 1975, 1977) or Monte Carlo procedures where bias was statistically generated by a computer (Merz & Grossen, 1968; Rudner, Getson & Knight, 1979). Although these computer studies have been able to control the amount of bias in test analysis, these studies have defined bias according to an arbitrary choice of model (the item characteristic curve is the one frequently used) and the data are of necessity artificial.

A further problem with detecting item bias is that measuring the bias in items against an internal criterion of the test as a whole is only logically valid to the extent that the test as a whole is considered to be less biased than the individual items. When attempting to control for ability differences, the methods do so with biased items used to measure the ability which is used to measure the bias in items. Thus, this whole circular process confounds ability and bias and is likely to be affected by the proportion of biased items. This problem of circularity is particularly important in minority testing where observed differences in test scores of one standard deviation have been found (Linn, 1973) and ability differences and differences due to bias are confounded to an unknown degree.

The present study extends previous research by having the realism

of an actual data set but in more controlled analysis situations. Males and females were chosen for study for several reasons, the most important of which are that:

1. The study was not designed to examine CONTENT bias against any particular group; it was designed to test which METHODS are functioning properly; and

2. The study COULD NOT answer the questions about the sufficiency of the methods if blacks and whites were used. This important design issue is discussed further later.

The conditions of the study, however, were chosen to have direct relevance to minority testing. The present study addresses several of the issues raised herein by:

1. Having items that are logically generated and evaluated a priori to be "biased" or "unbiased";

2. Analyzing tests composed of specified proportions of bias rather than having the amount of bias unknown; and

3. Selecting samples so that observed test score differences are one standard deviation apart (to simulate black/white differences), but when these observed differences are a result of known amounts of combinations of ability differences and differences due to bias.

The study compares the efficacy of four methods in identifying bias items in each of the conditions of 2 and 3 above. The four methods chosen for study were the transformed item difficulty approach,

the chi-square procedure, the three parameter item characteristic curve procedure, and the one parameter item characteristic curve procedure. These were chosen for study since they showed promise from previous studies.

In addition to the theoretical question, an important practical question was being addressed as the methods differ widely in cost, sample size required, and sophistication required to understand and implement the method. The three parameter method is very costly, requires very large sample sizes, and sophisticated background knowledge. On the other end of the continuum is the transformed item difficulty procedure. This procedure can easily be implemented, requires a much smaller sample size, and requires less mathematical sophistication to understand.

Thus, the study is designed to determine which method of detecting biased items is best and under what conditions.

REVIEW OF ITEM BIAS METHODS

Example of a biased item. Suppose, as part of a general information test, a Canadian is asked to answer the question: "How many senators are there in the U.S. Congress?" This question would likely be regarded as biased against Canadians because, according to one popular definition of bias, essentially equal ability Canadians and Americans would have an unequal chance of getting this item correct.

If we embedded this item in a larger test of general information, it might be identified statistically by the various procedures described below.

Transformed item difficulty. In this approach, an item is considered biased if for a given group it is relatively more difficult than other items on a test. The first step in this procedure involves calculating the item difficulty or p-value (proportion of subjects getting the item correct) for each of the two groups on each of the items. The p-values are then transformed into normal deviate Z values; i.e., Z is the tabled value having proportion (1-p) of the normal distribution below it. Then a delta value ($\Delta = 4Z + 13$) is calculated from the tabled Z to eliminate negative Z values, so that a large delta value indicates a difficult item. The pairs of transformed delta values (one pair for each item) are plotted on a bivariate graph, each pair being represented by a point on the graph.

The plot of these points appears as an ellipse extending from lower left to upper right. In order to identify the biased items, it is necessary to determine the major axis of the ellipse and the distance of the items from that axis. Those items that are relatively more difficult for one group than another fall at some distance from the axis and are identified as biased. The equation to be used for the major axis of the ellipse is given by Angoff and Ford (1973, p. 98). If the major axis is denoted by $Y = AX + B$, then the formula for the

perpendicular distance Di, of each point i, in the plot to the line is given as:

$$Di = \frac{AX_i - Y_i + B}{\sqrt{A^2 + 1}}$$

This perpendicular distance is a measure of the relative difficulty of an item and thus is a measure of the item's bias.

Chi-square. Two chi-square type procedures were calculated. The first considers only proportion correct and is therefore not a true chi-square statistic. According to Scheuneman (1975), "An item is unbiased if, for all individuals having the same score on a homogeneous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered" (p. 2).

The first part of the chi-square analysis involves establishing ability intervals for each item. This is accomplished with data from two sets of distributions. First, standard frequency distributions of total test scores are plotted separately for each group, and second, bivariate distributions of the number of correct responses to each test item by group. Scheuneman (1979) notes that at least three ability levels (calculated from total test scores) must be used for each item and there is little to be gained by having more than five intervals. Therefore, the study attempts to use five ability levels for each item

unless the following two criteria cannot be met (in which case fewer than five will be used). First, each score interval must contain five or more correct responses from each group. Second, the probability of a correct response within a given interval must be between 0 and 1.

The second part of the analysis involves the calculation of the chi-square value. The degrees of freedom for the test are reduced to (a-2)(b-1), where a is the number of ability levels and b is the number of groups. The degrees of freedom are (a-2) for the ability dimension because both the ability level of the examinee and the probability of a correct response must be estimated from the sample data. In addition, the formula for the expected cell frequencies in this procedure is different from the standard chi-square procedure. Algebraically, it is:

$$E_{xy} = \frac{A.y}{B.y} \quad C_{xy}$$

where A.y is the number of examinees in ability level y responding correctly; B.y is the total number of examinees in ability level y; C_{xy} is the total number of examinees in ability level X and group Y. The value of chi-square is calculated, a large chi-square indicating much bias.

The second chi-square type procedure follows the same logic but includes both correct and incorrect responses. It is described fully in Ironson (1982) and will be referred to in this report as the full chi-square.

Three parameter item characteristic curve: An item characteristic curve (ICC) specifies the relationship between the probability of an examinee answering an item correctly and his ability level (Birnbaum, 1968; Lord & Novick, 1968). The equation for the three parameter logistic model is given by Birnbaum (1968):

$$P(U_g=1/\theta_i)=c_g+(1-c_g)[1+\exp(+7a_g(\theta_i-b_g)]^{-1}$$

where $(U_g=1/\theta)$ is the probability of a correct response to item g given an examinee of ability level θ_i; a_g is an item discrimination index; b_g is an item difficulty index; c_g is a pseudo-guessing parameter. The curve for each item is determined from three parameters (ag, bg, and cg) that are estimated by the LOGIST procedure (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976). This is done separately for each of the two groups. However, in the present study, the test is composed of free response items so that only the difficulty and discrimination parameters need to be estimated. Thus, in the formula above, the cg parameter becomes zero. An unbiased item is one whose parameter values and item characteristic curves are the same for different ethnic groups, after equating (putting the parameter values on the same scale).

One parameter item characteristic curve: In this model, also referred to as the Rasch model, the probability of a correct response is a function of an examinee's ability and only one item parameter--difficulty. Formulas are given in Wright (1977). Items can first be tested for fit

to the model (Wright & Mead, 1977). Bias is measured by a shift in difficulty value for an item in the two groups (Draba, 1978; Durovic, 1975; Wright, Mead, & Draba, 1976). A $\underline{t}$ statistic is used for this purpose:

$$\underline{t} = \frac{d_{1i} - d_{2i}}{\sqrt{(Se_1)^2 + (Se_2)^2}}$$

where d_{1i} is the difficulty estimate for item i in group 1 and (Se_1) is the standard error for group 1. If $\underline{t}$ is large, an item is relatively more difficult for one group and is thus biased.

PROCEDURES

Sample. The sample used for the present study consisted of 533 male and 590 female undergraduates. The procedure requiring the largest number of examinees was the three parameter model using LOGIST. The sample size of over 500 is sufficiently above the minimum required since the test is long (60 to 100 items) and use of free response data means that the "c" parameter does not have to be estimated (Hulin, Lissak, & Drasgow, 1981).

Males and females were chosen as the most appropriate samples to use in this study (even though the most pervasive questions of bias deal with testing blacks and whites) for several important reasons:

1. First, the study was not designed to examine CONTENT bias against any particular group; it was designed to test which METHODS are functioning properly.

2. Second, and most importantly, the study COULD NOT answer the questions about the sufficiency of the methods if blacks and whites were used. The cultural groups chosen were irrelevant except for two important stipulations:

A. Groups chosen had to be UNCONTROVERSIALLY roughly equal in ability; and

B. Both groups had to be able to agree on which items are biased and unbiased. Anyone who has worked with blacks knows this simply does not hold. For example, some blacks may feel that all items reflect white middle class culture, perhaps justifiably so. Furthermore, no one knows how much of the observed difference between blacks and whites is due to ability and how much is due to bias.

Without agreement on which items are biased and unbiased, it would be difficult to separate differences due to bias and differences due to ability.

Choosing groups roughly equal in ability[1] and who could agree on biased and unbiased items enables us to:

---

[1]It was thought that males/females at this university would be roughly of equal ability. This assumption was checked by comparing observed distributions of males and females on the 60 unbiased items, and will be discussed later in the Results section.

A. Circumvent problems associated with artifacts noted by Hunter (1975) that are due to distributions of unequal ability;

B. To examine the contribution of ability differences (measured by unbiased items) and differences due to biased items in producing observed total score variattions. Furthermore, we can observe how this affects the detection of biased items;

C. Simulate observed black/white differences but in a situation where we can tell what is due to an ability difference and what difference is due to bias.

Research instrument development: The research instrument developed for use in this study was designed to measure general information. As a starting point, the unbiased items were constructed parallel to those on the general information section of the WAIS. For example, instead of the question, "Who wrote Faust?" one question was "Who wrote Catcher in the Rye?" For the items intended to be biased, samples of males and females were asked to generate items with these directions: Given a male and female of equal ability on general information, give examples of items that a male would have a greater probability of getting right.

A preliminary instrument was generated consisting of 150 items (57 biased and 93 unbiased). A sample of 32 males and 41 females was asked to evaluate each of the items for bias by rating each item on a 5-point scale (from unbiased = 1 to biased = 5) where bias was defined as above;

i.e., given equal ability, males have a greater probability of getting the item right. A different sample (37 males, 37 females) was asked to answer the 150 questions, so that information on the difficulty and item to total correlations could be obtained.

The reliability of the 150-item test in the combined sample was .94 (Cronbach's Alpha).

Items were defined as biased if more than 55% of the combined male and female (N = 74) group gave it a 4 or 5 and there was no significant difference between the male and female sample rating. (Low bias included 55-75%, medium bias 75-85%, and high bias 85-95% rating it a 4 or 5.) In addition, items were dropped if they had "p" values of greater than .95 or less than .05, or point biserials less than .15 (combined sample). From the items surviving the above, a final instrument comprised of 110 items (65 unbiased, 45 biased) was administered.

Characteristics of data sets. Each of 533 males and 590 females took the 110-item research instrument. Six different data sets were constructed from the initial data base so that (1) the percentage of biased items could be varied and (2) the observed differences in unbiased score distributions could be set to one standard deviation apart. The purpose of this manipulation was to mirror what is often found in black/white scores on tests, but in a controlled situation where the source of the difference (ability or bias) would be known.

1. Percentage of biased items. The data were analyzed with 0%

biased items (60 items; all unbiased), 23% biased items (78 items; 60 items unbiased, 18 items biased[2]) and 40% biased (100 items; 60 unbiased, 40 biased). These three amounts of bias were chosen because they approximate what has been found in empirical studies. For example, Scheuneman (1975, 1977) found 14% to be biased against blacks; Ironson and Subkoviak (1979) found 24% biased against blacks; Scheuneman (1976) found 35% to be biased across several groups; and Rudner (1977) found 56% biased against hearing-impaired subjects.

2. Observed differences in score distributions of males and females. It was felt that males and females at this university would be of roughly equal ability. Therefore, the first three of the six data sets would be generated solely by changing the proportion of biased items. Data sets four, five, and six would be generated by selecting out high ability females and low ability males so that unbiased score differences (on the unbiased items) would be set to one standard deviation apart.

The assumption of equal ability males and females was checked by examining the distribution of males and females on the 60 unbiased items. The males were approximately one-half of a standard deviation above the females ($\bar{X}$ males = 33.11, S = 10.22; $\bar{X}$ females = 28.75,

---

[2]Eighteen items were chosen for bias so that items would cover a "p" value range (easy, medium, hard) and items would cover a bias range (low, medium, high).

SD = 10.41). Because of this unexpected difference, another measure of ability was obtained--grade point average. On this measure, females were about one quarter of a standard deviation above males ($\bar{X}$ males = 2.58, S = .67; $\bar{X}$ females = 2.77, S = .65). Since the discrepancies were in opposite directions, their abilities were seen as roughly equivalent.

In order to create the unequal ability groups with an observed one standard deviation difference between males and females, the following procedure was used. The desired difference between males and females was targeted at one standard deviation, or about 10 points (with a standard deviation of about 10 points). This would mean moving the male mean up about 2-3 points and moving the female mean down about 2-3 points, while maintaining the shapes of the respective distributions. Knowing the desired mean and standard deviation of the female distribution, we calculated the proportions of females at each score level which would give this. Then females were randomly sampled from each score level to achieve numbers that reflect those proportions. We then repeated the same procedure for males. This resulted in a reduced sample (N = 909; 433 males, 476 females) with a one standard deviation difference ($\bar{X}$ males = 35.64, S = 9.14; $\bar{X}$ females = 26.20, S = 9.31) on the 60 unbiased items.

Six data sets used. Thus, the first three data sets were composed of roughly equal ability males and females (using the original sample) where the instrument was analyzed with 0%, 23%, or 40% biased items. Data sets four, five, and six were composed of a reduced sample of males and females manipulated to be one standard deviation apart where the instrument was again analyzed with 0%, 23%, and 40% biased items.

Table 1 summarizes the characteristics of the six data sets used in the study. They can be placed in a continuum. Set I is the small difference in ability, no bias. Set VI is the large ability difference, large bias amount. The various sets in between vary in relation to the bias amount and the ability differences present.

Table 1. Characteristics of Data Sets Used in This Study.

Data Set Abbreviation	Size	M/F Diff. on 60 Unbiased Items	% Biased Items	Number of Unbiased Items to Total Items
I. S60	533M 590F	1/2 SD	0	(0/60)
II. S78	"	"	23	(18/78)
III. S100	"	"	40	(40/100)
IV. L60	433M 476F	1 SD	0	(0/60)
V. L78	"	"	23	(18/78)
VI. L100	"	"	40	(40/100)

M = males; F = females; SD = standard deviation.

RESULTS

For each of the six conditions (three proportions of biased items with or without manipulation to achieve a one standard deviation difference on the unbiased items), five item bias techniques were calculated and are described below.

Bias Methods Used in the Analysis

1. Transformed Item Difficulty (TID). The distance Di from the major axis of the ellipse was computed for each item and used as the measure of bias. The sign indicating direction of bias was maintained. A positive sign indicates an item that is relatively more difficult for females.
2. One parameter item characteristic curve (1ICC): The difficulty parameter for each item in each group was estimated by BICAL. The t statistic as described previously was used. A positive t represents an item biased against females.
3. Scheuneman chi-square (SCHI). Scheuneman's chi-square, which considers only proportion correct for each ability level, was obtained for each item. Each item was given a sign according to the direction of the difference of p value within ability levels. A positive sign indicates bias against females.
$. Full chi-square (FCHI). The full chi-square, which includes both correct and incorrect proportions for each ability level, was

obtained for each item. As in Scheuneman's chi-square, each item was given a sign according to the direction of the difference of p values within ability levels. A positive sign indicates bias against females.

5. Three parameter item characteristic curve area (AREA). The ICC estimated separately for each group by the LOGIST program. After linear equating, the area between the ICC for females and males was computed by the formula given in Rudner, Getson, and Knight (1979). A positive sign attached indicates bias against females.

The means and standard deviations of each of the bias methods are given in Table 2. The most striking result of that table is a general trend for the bias indices and their variability to be markedly less in the conditions (I and IV) with no biased items. (Caution must be exercised in interpreting the chi-square indices here because only items that could be divided into five ability intervals are included in this table. As is noted at the bottom of the table, this resulted in a considerable loss of items particularly when there was a high percentage of biased items and a large ability difference. Of additional note in interpreting the table is to keep in mind that the signed TID always sums to zero.) Of additional interest is a finding of close similarity in means and standard deviations across small and large ability differences (contrasting I and IV, II and V, III and VI). Thus, the major contributor to differences appears to be whether there are biased items or not.

Table 2. Means and Standard Deviations of Bias Methods.

Data Set	TID	1ICC	SCHI Unsigned	SCHI Signed	FCHI Unsigned	FCHI Signed	AREA Signed	AREA Unsigned
I-S60	.000	-.036	6.48	+.047	12.907	.605	.064	.379
	(.626)	(2.935)	(6.58)	(9.097)	(11.807)	(17.189)	(.376)	(.218)
II-S78	.000	.262	14.322	+3.508	27.656	6.602	.113	.624
	(1.10)	(4.73)	(18.688)	(23.257)	(33.167)	(42.661)	(.678)	(.387)
III-S100	.000	.279	15.676	+2.378	30.602	4.502	.136	.609
	(1.16)	(5.20)	(20.168)	(25.287)	(34.124)	(45.44)	(.693)	(.411)
IV-L60	.000	-.006	5.586	+1.975	11.500	4.241	.046	.382
	(.670)	(2.75)	(5.482)	(7.448)	(10.34)	(14.583)	(.386)	(.205)
V-L78	.000	.380	9.328	5.206	18.082	+10.115	.292	.597
	(1.07)	(3.84)	(12.156)	(14.308)	(22.318)	(26.73)	(.702)	(.492)
VI-L100	.000	.280	6.41	+.460	13.309	2.49	.203	.583
	(1.08)	(4.197)	(5.249)	(8.01)	(10.705)	(16.57)	(.691)	(.458)

CHI had these sample sizes:
(Number of items that could not be evaluated is in parenthesis in five intervals)

I - 54(6) II - 64 (14)
III - 79(21) IV - 52(8)
V - 42(26) VI - 51(49)

Table 3. Intercorrelations Among Signed Bias Indices.

	TID	1ICC	*SCHI	*FCHI		TID	1ICC	*SCHI	*FCHI
	(I. S60)					(IV. L60)			
1ICC	.99					.99			
SCHI	.89	.90				.82	.83		
FCHI	.93	.94	.96			.87	.87	.96	
AREA	.95	.95	.92	.95		.82	.82	.91	.94
	(II. S78)					(V. L78)			
1ICC	.98					.94			
SCHI	.90	.90				.87	.86		
FCHI	.93	.92	.98			.89	.88	.98	
AREA	.92	.94	.92	.95		.87	.91	.91	.93
	(III. S100)					(VI. L100)			
1ICC	.99					.99			
SCHI	.85	.84				.38	.40		
FCHI	.90	.89	.98			.41	.43	.97	
AREA	.92	.92	.90	.93		.82	.82	.44	.47

* Sample sizes for the CHI procedures are:

I - 54/60; II - 64/78; III - 79/100; IV - -2/60; V - 42/78; VI - 51/100.

This is because items that could not be evaluated in five intervals were dropped.

Agreement Among Bias Methods

Table 3 gives the intercorrelations among the bias methods for the six conditions of the study. For the first three conditions (small ability difference), the intercorrelations are for the most part in the 90s. The percent of biased items does not appear to make a difference. The high agreement among methods is also apparent for conditions IV and V. The lower agreement for the chi-square techniques in condition VI (large ability difference) may be due to the large number of items which could not be evaluated using five intervals.

There unfortunately is no easy way of putting items that have been evaluated with a chi-square using a different number of intervals back on the same scale. The unsigned significance or "p" value could be used. These would change the correlations between SCHI and TID, 1ICC and AREA to .25, .26, .21; and between FCHI and TID, 1ICC and AREA to .27, .27, .22. These are based on an N of 96. (Four items could not even be evaluated with only two ability levels.) That the correlations are lower using "p" values is not surprising. In the first place, correlations between the chi-square value and the "p" value of the significance test is only roughly .6-.8. Secondly, an earlier study (Rudner, 1977) also found the "p" values did not function well.

In general, then, Table 3 shows excellent agreement among the methods except for the sixth condition with the chi-square techniques. The agreement does not appear to be affected by either the percent of

biased items or the ability difference, except insofar as a large ability difference affects the computation of the chi-square.

Agreement Between Subjective Methods of Bias and Statistical Bias Methods

The correlations of bias indices with judged bias are presented in Table 4. For the conditions of small ability differences--I, II, and III, the correlation between the indices and judged bias is moderate (.7-.8) when there are biased items, and low (.3-.4) when there are not. For the large ability difference conditions IV and V, the same pattern repeats: high correlations when there are biased items (V), low correlations where there are no biased utems. In examining condition (VI) which has a preponderance of biased items (40%), TID and 1ICC have the highest correlation with judged bias. The lower correlations of the chi-square techniques may be due to the loss of 49 items that could not be evaluated with 5 intervals. Using unsigned "p" values instead did not alter the correlations appreciably (.29, .35). The reason for the lower AREA correlation was not immediately apparent. One possibility that was explored and rejected was that lack of unidimensionality in condition IV may have harmed the AREA measure (see the section of factor analysis results). Both the percent of variance and ratio of first to second eigenvalues were extremely stable across the six conditions.

Table 4. Correlation of Signed Bias Indices with Judged Bias.

	I S60	II S78	III S100	IV L60	V L78	VI L100
TID	.38	.84	.87	.31	.78	.83
1ICC	.37	.83	.86	.32	.75	.82
*SCHI	.47	.78	.66	.46	.80	.24
*FCHI	.45	.80	.72	.44	.82	.27
AREA	.42	.84	.77	.43	.83	.67

* CHI sample sizes are reduced; see Table 3, p. 21.

Agreement Between Statistical Methods and Biased/Unbiased Classification

Table 5 applies cutoff values found in the literature to the identification of the biased items. TID is biased if it is greater than 1.5 (Strassberg-Rosenberg & Donlon, 1975), 1ICC is biased if it is greater than 2.4 (Draba, 1979), CHI is biased if it is significant at the .05 level, and AREA is biased if it is greater than .70 (Merz & Grossen, 1978).

Using these cutoffs, there are several trends that are apparent in Table 5. The first is that the cutoffs work best with a smaller proportion of biased items (18 vs. 40). The cutoffs also work better with the small ability difference data sets (II and III). For the large ability group differences, the 1ICC and FCHI seem to work best. Finally, the TID cutoff appears to be too low, because a lot of biased items are missed.

Two additional points are essential to note in interpreting this table. The first is that the cutoffs were applied in only one direction. This means that, if the cutoff for TID was +1.5, then an item with a TID of -1.6 was not considered biased. Similarly, the "p" values were calculated for SCHI and FCHI, a sign was attached indicating the direction. Again, if the sign was in the opposite direction even if the "p" value was small, the item was not declared as biased.

Table 5. Identification if Biased Items Using Selected Cutoffs for Statistical Procedures.

Data Set	Method				
	TID (> 1.5)	1ICC (> 2.4)	SCHI (p < .05)	FCHI (p < .05)	AREA (> .70)
II-18biased	11	18	13	18	16
V-18biased	9	17	16	17	16
III-40biased	26	33	26	36	17
VI-40biased	7	29	10	25	16

A second additional point regards the number of items incorrectly identified as biased (by statistical procedures) when they are unbiased (by subjective judgment). For most of the conditions, the number was less than 5 with one notable exception: For the 60 unbiased items (I and IV) the 1ICC incorrectly identified 12 and 13 items, respectively, as biased. The averages over all 6 conditions out of a possible total of 60 unbiased items were: TID-0, 1ICC-5.2, SCHI-2.3, FCHI-6.5, AREA-2. While these numbers may seem fairly low, if one disregards sign, the numbers increase dramatically (see paragraph above). That is, many items were identified as biased against males when they were intended to be unbiased items.

Psychometric Properties of the Tests

Tables 6 through 9 describe the effects of varying the ability differences and proportion of biased items on the psychometric properties of the test. Descriptive information is provided in Table 6. Tables 7 through 9 provide information on reliability, unidimensionality, and validity, respectively.

Table 6 shows the effect of ability differences and bias amount on the observed means and standard deviations. The difference in ability magnifies the difference between males and females 1.65 times. More interesting, however, is the proportion of biased items: increasing it to 23% magnifies the difference 1.6 times; increasing it to 40% magnifies the difference 2 times. Thus, both ability

Table 6. Observed Means and Standard Deviations for Six Conditions.

	Small Ability Difference Males(533)	Females(590)		Large Ability Difference (Males(433)	Females(476)
I.S60	33.11(10.22)	28.75(10.41)	IV.L60	35.64(9.14)	26.20(9.31)
II.S78	45.11(11.91)	39.36(12.54)	V.L78	48.03(10.42)	32.42(11.31)
III.100	61.74(14.91)	45.45(16.15)	VI.L100	65.29(12.94)	41.96(14.78)

Table 7. Reliabilities of Tests Composed of Varying Amounts of Bias.

	Males	Females
(Small Ability Difference Groups)		
I.S60 (0% biased)	.8981	.9007
II.S78 (20% biased)	.9036	.9123
III.S100 (40% biased)	.9232	.9315
(Large Ability Difference Groups)		
I.L60 (0% biased)	.8726	.8761
II.L78 (20% biased)	.8746	.8927
III.L100(40% biased)	.8993	.9188

Table 8. Factor Analysis Results (Males & Females Combined).

Condition	% Variance Explained by First Factor	Ratio of First to Second Eigenvalue
N = 1123:		
I. S60	15.8	4.18
II. S78	14.6	3.26
III. S100	15.7	2.86
N = 909:		
IV. L60	15.4	3.94
V. L78	15.1	3.83
VI. L100	17.0	3.56

Table 9. Validities of Tests Composed of Varying Amounts of Bias (correlation with GPA).

	Male	Female
Unbiased items	.21	.32
Biased items	.04	.13
Small Ability Differences:		
I. 0% biased	.21	.32
II. 20% biased	.18	.29
III. 40% biased	.15	.25
Large Ability Differences:	(N=354)	(N=406)
IV. 0% biased	.16	.35
V. 20% biased	.13	.32
VI. 40% biased	.10	.27

Note 1: Item-total correlations on M&F combined:

Mean = .23 (s.d. = .12) for unbiased items
Mean = .41 (s.d. = .11) for biased items.

Note 2: On the combined male and female sample the validity of the unbiased items was .23; the validity of the biased items was -.02.

differences and the proportion of biased items had a pronounced effect on observed differences.

Table 7 presents the reliabilities for the six conditions. The reliabilities are all high and do not seem to be much affected by either ability difference, sex, or proportion of biased items. Reliabilities are slightly higher for longer tests, which is what one would expect.

Table 8 presents the principal components factor analysis results in order to investigate unidimensionality. The test appears to be marginally unidimensional as indicated by high reliability, and ratios of first to second eigenvalue of about 3. The percent of variance explained by the first factor was remarkably stable across all six conditions of the study. This was particularly surprising because one would expect the conditions with biased items on them to be multidimensional. In fact, one hypothesis for the poorer performance of the AREA measure in condition six was that it lacked the unidimensionality required for a latent trait analysis. However, it was no more nor less unidimensional than the other conditions. Furthermore, Reckase (1979) recommends 15-20% variance explained by the first factor for a latent trait analysis.

Table 9 presents the validity data (correlation of research instrument with GPA). Several findings are evident from the table. This test appeared to be more valid for females than for males.

Secondly, for both men and women, the unbiased items were more valid than the biased items. To check to see whether this result may have been due to the unbiased items simply being better items, the item-total correlations were examined. In fact, the biased items had higher correlations. Third, as the proportion of biased items went up, the test validity went down for both males and females (but not by a large amount). To hope that item bias studies will eliminate test bias is being overly optimistic.

CONCLUSION

The purpose of this study was to determine the effect that amount of bias and amount of ability difference would have on (1) agreement among the statistical methods, (2) agreement between the statistical methods and the subjective judgments of bias, (3) the psychometric properties of the tests such as reliability and validity.

The following is a summary of the major results:

1. The agreement among the methods is very high except for data set VI (large bias and large ability difference) for the chi-square techniques. This is very likely an artifact of loss of items due to an inability to use five intervals in calculating the chi-square. A procedure for getting chi-squares back on one scale needs to be developed. Otherwise the agreement was high, especially compared to other studies in the published literature (Burrill, 1982).
2. The correlation between statistical indices and judges bias was moderate when there are no biased items (which one would expect due to restriction of range), very high when there are biased items (II, III, and V) and not as good when there are both large ability differences and the amount of bias is large (VI). In the latter case, TID and 1ICC performed best.
3. Cutoffs for all procedures except TID appear to be working moderately well when only 23% of the items are biased and the direction

of bias is taken into account. The TID procedure underidentifies biased items. All methods would overidentify items as biased if the sign were not taken into account (due to the interaction nature of the methods). Studies measuring bias in both directions most probably overestimate the amount of bias with these cutoffs.

When the amount of bias goes up to 40%, all of the procedures miss a fair number of items especially when there are large ability differences. Despite this, the 1ICC and FCHI do the best job of identifying these items.

4. The presence of biased items increases the mean score differences between males and females, does not seem to affect reliability, and does decrease validity (although not by a large amount).

The major overall result is that the methods seem to be working well except for the last condition. It should be noted, however, that the last condition is rather extreme. Although there is a one standard deviation difference on the unbiased items, there is approximately a two standard deviation difference in total score.

This research points out several areas for further work. Effort needs to be directed toward improving cutoffs for the procedures. Sampling distributions need to be developed. A method needs to be developed for putting chi-squares calculated on different numbers of intervals back on the same scale.

It was felt that the procedures attempting to control for ability would work better than those not controlling for ability when the ability difference was varied. This turned out not to be the case. The dimensionality hypothesis was rejected as a cause for this. It may be, however, that the area and chi-square procedures would work better with larger sample sizes. Expected frequencies within each interval would then be adequate.

Finally, the practitioner may be comforted in knowing that, if the ability differences are 1/2 a standard deviation or less on the unbiased items or 23% less of the items are biased, all of the methods agree fairly well.

BIBLIOGRAPHY

Angoff, W. H., & Ford, S. F. Item-race interavtion on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick: Statistical theories of mental test scores. Reading, MA: Addison-Wesley, Chapters 17 through 20.

Cardall, C., & Coffman, W. R. A method for comparing performance of different groups on the items in a test (RM 64-61). Princeton, NJ: Educational Testing Service, 1964.

Cleary, T. A., & Hilton, T. L. An investigation into item bias. Educational and Psychological Measurement, 1968, 8, 61-75.

Draba, R. E. The Rasch model and legal criteria of a reasonable classification. Unpublished doctoral dissertation, University of Chicago, 1978.

Durovic, J. An objective definition of test bias. Paper presented at the Annual Meeting of the Northeastern Educational Research Association, Ellenville, NY, 1975.

Green, D. R. Reducing bias in achievement tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1976.

Green, D. R., & Draper, J. F. Exploratory studies of bias in achievement tests. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu, September 1972.

Hulin, C. L., Lissak, R., & Drasgow, F. Effect of sample size and test length on IRT parameter estimation. Paper presented at the Annual Meeting of the American Psychological Association, Los Angeles, August 1981.

Hunter, J. E. A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias. Maryland, December 1975.

Ironson, G. H. A comparative study of several methods of assessing item bias. Unpublished doctoral dissertation, University of Wisconsin, 1977.

_____. A comparative analysis of several methods of assessing item bias. Paper presented at the American Educational Research Association Convention, Toronto, March 1978.

_____. Chi-square and latent trait approaches to measuring item bias. In Berk, R. (ed.): Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press, 1982, pp. 117-160.

Ironson, G. H., & Subkoviak, M. J. A comparison of several methjos of assessing item bias. Journal of Educational Measurement, 1979, 16(4), 209-225.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

Lord, F. M. A study of item bias using item characteristic curve theory. Preliminary report. Princeton, NJ: Educational Testing Service, 1977.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Boston, MA: Addison-Wesley, 1968.

Merz, W. R. Factor analysis as a technique in analyzing item bias. Paper presented at the Annual Meeting of the California Education Research Association, Los Angeles, 1973.

_____. Estimating bias in test items utilizing principal components analysis and the general linear solution. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976.

_____. Test fairness and test bias: A review of procedures. In M. Wargo & D. R. Green, eds.: Achievement testing of disadvantages and minority students for educational program evaluation. Monterey, CA: McGraw-Hill, 1978.

Merz, W. R., & Grossen, H. E. An empirical investigation of six methods for examining test item bias. Sacramento, CA: Foundation of California State University, 1978.

Nungester, R. An empirical examination of three models of item bias. Unpublished doctoral dissertation, Florida State University, 1977.

Ozenne, D. G., van Gelder, N. C., & Cohen, A. J. Emergency School Aid Act (ESAA) National Evaluation, Achievement Test Standardization. Santa Monica, CA: Systems Development Corporation, 1974.

Petersen, N. S. Bias in the selection rule: Bias in the test. Paper presented at the Third International Symposium on Educational Testing, University of Leyden (Netherlands), June 1977.

Rudner, L. M. Item and format bias and appropriateness. Unpublished doctoral dissertation, Catholic University of America, 1977.

Rudner, L. M., & Convey, J. J. An evaluation of select approaches for biased item identification. Paper presented at the Annual Meeting of the American Educational Research Association. Toronto, March 1978.

Rudner, L. M., Getson, P. R., & Knight, D. L. A Monte Carlo comparison of five biased item detection techniques. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1979.

_____. Biased item detection techniques. Journal of Educational Statistics, 1980, 5(3),

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC, April 1975.

_____. A procedure for evaluating item bias in the absence of an outside criterion. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1976.

_____. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Strasberg-Rosenberg, B., & Donlon, T. F. Context influences on sex differences in performance and aptitude tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, DC, 1975.

Veale, J. R., & Foreman, D. I. Cultural variation in criterion referenced tests: A global item analysis. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1976.

Williams, R. L. Black pride, academic relevance, and individual achievement. The Counseling Psychologist, 1970, 2, 18-22.

_____. Abuses and misuses of testing black children. The Counseling Psychologist, 1971, 2, 62-73.

Wood, R. L., & Lord, F. M. A User's Guide to LOGIST. Research memorandum. Princeton, NJ: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research memorandum. Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., & Mead, R. J. The use of measurement models in the definition of social science variables. Army Research Institute Technival Report, June 1977.

Wright, B. D., Mead, R. J., & Draba, R. Detecting and correcting test item bias with a logistic response model. Research memorandum No. 22, Statistical Laboratory, Department of Education, University of Chicago, 1976.